HERAUSFORDERUNGEN BEI OPEN SOURCE INTELLIGENCE (OSINT)

Lennart Karsten,
Hochschule für Angewandte Wissenschaften Hamburg, Dept. Informatik,
Berliner Tor 7
20099 Hamburg, Germany
lennart.karsten@haw-hamburg.de

ABSTRACT

Durch den massiven Anstieg in der Nutzung von Social Media Plattformen, wie Twitter oder Facebook, hat dessen Wichtigkeit für die Gewinnung von „Business intelligence“ Informationen in den vergangenen Jahren stark zugenommen.
Die Erhebung von Daten aus solchen, meist frei zugänglichen Quellen ist deutlich kostengünstiger als die Erhebung aus klassischen Quellen. Dies hat dazu geführt, dass Open Source Intelligence (OSINT) mittlerweile die wichtigste Quelle zu Gewinnung von „Business intelligence“ Informationen ist.
Im Gegensatz zu klassischen Aggregierungsmethoden ist die Beschaffung einer ausreichenden Menge an Daten beim OSINT kein Problem. Schwierigkeiten ergeben sich durch große Mengen an Daten, die teilweise widersprüchlich oder falsch sein können.
Die Ausarbeitung befasst sich mit der Beschreibung solcher Problemstellungen und nennt verschiedene Lösungsansätze.

1. EINLEITUNG

Das Finden von Informationen zur Entscheidungsfindung war noch bis vor ein paar Jahren eine aufwendige Aufgabe. Um ein umfassendes Bild über eine Personengruppe oder eine einzelne Person zu erlangen mussten eine Vielzahl von Quellen ausgewertet werden. Der Vorgang war zeitintensiv und somit teuer.
Durch die alltäglichen Nutzung des Internets durch Millionen von Personen hat sich das Bild gewandelt. Es ist mittlerweile möglich Daten massenhaft zu speichern und auszuwerten. Die Qualität der gesammelten Daten hat mit dieser neuen Methode jedoch massiv abgenommen. Dies ergibt sich daraus, dass jeder Benutzer beliebige Daten im Netz über sich streuen kann. Um Die Daten auswerten zu können müssen diese vergleichbar gemacht werden. Außerdem muss bedacht werden, dass Ergebnisse aus einer derartigen elektronischen Auswertung niemals zu 100% korrekt sein können.

2. VERGLEICHBARE ARBEITEN

[1] „Challenges in Open Source Intelligence“
Beschreibt allgemeine Schwierigkeiten bei OSINT. Der Schwerpunkt liegt bei dieser Arbeit auf der Filterung von erhobenen Daten um diese nutzbar zu machen.

[2] „Challenges to Automated Allegory Resolution in Open Source Intelligence“
Die Arbeit befasst sich insbesondere mit Problemen bei der Auswertung von Sprache an konkreten Beispielen.

[3] „Development of a Hybrid Decision Support System for intelligence analysis“
Im Gegensatz zu den übrigen Arbeiten befasst diese sich nicht mit dem Versuch Analyse nach Möglichkeit zu automatisieren. Stattdessen wird OSINT als Werkzeug, welches dem Analysten Hilfestellungen liefert begriffen.

[4] „Data consolidation solution for internal security needs“
In der Arbeit wird der Einsatz in Indien beschrieben. Der Fokus liegt in dem Versuch alltägliche Abläufe in der Gesellschaft zu optimieren.

3. DATENSAMMLUNG

Das Sammeln von Daten für die Durchführung von OSINT unterscheidet sich maßgeblich im Vergleich zu klassischen Verfahren. Der folgenden Abschnitte beleuchten diese Unterschiede, da sie für die Art der Nutzung von Bedeutung sind.

3.1. klassische Datensammlung

Bei der klassische Gewinnung von Daten wird die Erhebung manuell durchgeführt. Es ist somit möglich, beliebige Quellen zu nutzen, z.B. Zeitungsartikel, handschriftliche Notizen, Tonaufnahmen oder mündliche Aussagen von Personen.
Die klassische Datensammlung geschieht mit dem Ziel, bestimmte Informationen heraus zu finden oder einen sehr speziellen Themenkomplex zu erschließen. Für diese Art der Datensammlung können jegliche Art von Daten genutzt werden, da eine manuelle Zusammenführung von Personen getätigt wird.

3.2. automatisierte Datensammlung

Beim OSINT werden die Daten zum größten Teil maschinell ausgewertet. Möglich ist eine solche Auswertung durch die gesellschaftlichen Veränderungen, die uns das WEB 2.0 gebracht hat.
Seit ein paar Jahren kommunizieren wir mit zunehmendem Maße schriftlich, digital und häufig öffentlich. OSINT macht sich diese Tatsache zu Nutzen. Es erfasst und wertet Daten aus Social Media Plattformen wie Twitter, Facebook, Blogs und anderen frei zugänglichen Quellen aus.
Da dies Auswertung automatisiert vorgenommen wird, ist die Verarbeitung von großen Mengen an Daten möglich.

3.3. Vergleich

Wie aus Tabelle 1 zu ersehen ist, eignet sich OSINT insbesondere dann, wenn viele Daten betroffen sind und keine hohe Korrektheit der Daten von Nöten ist.

Tabelle 1: *Vergleich von klassischer und digitaler Informationsgewinnung mittels OSINT*

Bereich	klassisch	OSINT
Aufwand der Erhebung (einmalig)	höher	niedriger
Skalierbarkeit der Erhebung	schlechter	besser
Aufwand der Auswertung (einmalig)	hoch	hoch
Skalierbarkeit der Auswertung	schlechter	besser
Qualität der Auswertung	höher	niedriger

4. DATENBEREINIGUNG

Gesammelte OSINT Daten enthalten eine Vielzahl von Fehlern. Hierzu gehören: Unterschiedliche Schreibweise von Wörtern, Dopplungen, sowie Widersprüche. Nach Prasad et al.[4] gibt es die folgenden Phasen.

4.1. Untersuchung

In dieser Phase werden allgemeine Fehler auf Basis von Querreferenzierung und Plausibilitätsüberprüfung gefunden und entfernt.

Tabelle 2: *Beispiel für den Prozess der Untersuchung*

vorher	nachher
Teststrasse	Teststrasse
Test Straß	Teststrasse
Teststr.	Teststrasse

4.2. Standardisierung

Anschließend werden die Daten in ein einheitliches Format gebracht um sie effizient durchsuchen zu können. Zudem werden Rechtschreibfehler korrigiert.

Tabelle 3: *Beispiel für den Prozess der Standardisierung*

vorher	nachher
Teststrasse	Teststraße
Teststrasse	Teststraße
Teststrasse	Teststraße

4.3. Beseitigung von Dopplungen

In der 3. Phase werden redundante Daten entfernt. Hierbei werden ähnliche Daten zunächst in einem Block zusammengefasst, wodurch die Effizienz der Abgleichung erheblich erhöht werden kann. Im 2. Schritt werden identische Daten pro Block zusammengefasst.

Tabelle 4: *Beispiel für Beseitigung von Dopplungen*

vorher	nachher
Teststraße	
Teststraße	Teststraße
Teststraße	

4.4. Domain Phase

Im letzten Schritt werden die gesammelten Daten auf Basis von Domainwissen gefiltert und Themenbereichen zugewiesen. Dies wird im folgenden Abschnitt näher beschrieben.

5. ZUORDNUNG DER OSINT DATEN

Die gesammelten OSINT Daten wurden bereinigt und normalisiert. Um sie effizient nutzen zu können ist es notwendig eine Zuordnung durchzuführen. Dies geschieht auf verschiedenen Ebenen. Diese sind:

- Thematische Zuordnung
- Geografische Position
- Namen (Named Entity Recognition (NER))
- Beziehungen
- Ereignisse

5.1. Thematische Zuordnung

Bedeutungen von Aussagen können in Bezug auf den Kontext in dem sie verwendet werden variieren. Um eine korrekte Aussage über die Bedeutung einer Aussage, oder eines Textes zu treffen, ist es somit notwendig, den Kontext zu bestimmen.
Bei vielen Themengebieten ist es zudem von Nöten, eine sprachabhängige Zuordnung durchzuführen. Dies ist notwendig, da in verschiedenen Ländern zu einem Themenkomplex ein anderer „Major consensus narrative“[1] vorherrschen kann.
Eine thematische Zuordnung kann durch den Einsatz von mehrsprachigen Stichwortlisten geschehen. Bei diesem Ansatz hat jedes bekannte Themengebiet Stichworte, die in einer bestimmten Reihenfolge zu einem gewichteten Ergebnis führen. Hierdurch wird die Wahrscheinlichkeit gemessen, zu der ein Text zu einem Thema gehört.
Alternativ gibt es die Möglichkeit die Themenzugehörigkeit mit Hilfe von linearer Algebra zu ermitteln.

5.2. Geografische Position

Um eine bessere Thematische Zuordnung durchführen zu können, ist es hilfreich den Aufenthaltsort des Verfassers einer Nachricht zu ermitteln. Dies ist allerdings weniger trivial, als man zunächst annehmen würde.

[1] Dieser wird von Bruce Sterling[5] als das beschrieben, was eine Personengruppe über ein Thema als wahr annimmt. Diese Wahrheit basiert auf dem erlebten und muss nicht zwangsläufig mit den Fakten übereinstimmen. Bsp.: Die gegenläufige Wahrnehmung Europas und Russlands zu dem Krieg in der Ukraine

Ein Grund hierfür ist, dass Orte häufig abgekürzt werden, oder unterschiedliche Schreibweisen haben. Hamburg kann z.B. HH abgekürzt werden und München wird im englischen „Munich“ geschrieben. Orte müssen daher, wie in 4.2 beschrieben, mit Hilfe einer Fuzzysuche angeglichen werden.
Erschwerend kommt hinzu, dass ein genannter Ort nicht zwangsläufig der Aufenthaltsort des Verfassers sein muss. In einigen Fällen kann der Aufenthaltsort aus Zusatzdaten ausgelesen werden. Dies könnte beispielsweise eine Twitter Nachricht mit Bild, das eine Georeferenzierung in ihren EXIF[2] Daten enthält, sein.
Wenn keine derartige Datenquelle verfügbar ist, muss der Aufenthaltsort aus dem Kontext gelesen werden. Dies erfordert jedoch abhängig von dem Thema ein entsprechend gute thematische Kenntnis. Verbessert werden kann eine derartige geografische Positionsbestimmung bei vielen Nachrichten zum gleichen Thema.

5.3. Namen (Named Entity Recognition (NER))

Um Texte und Aussagen einer Person zuordnen zu können ist es notwendig Namen in Texten zu erkennen und diese als Identität wieder zu erkennen. Dies wird „Named Entity Recognition (NER)“ genannt. Bei der NER gibt es zwei wesentliche Vorgehensweisen.
Einerseits ist es Möglich Namen anhand von bestehenden Listen zu erkennen und zuzuordnen.
Die andere Möglichkeit ist, Namen anhand von Mustern auszumachen. Z.B. [Anrede] [groß geschrieben] [groß geschrieben] oder [Title] von [Ort]. Bei diesem vorgehen kann eine ungefähre Genauigkeit von 80 Prozent[1] erreicht werden. Sprachen, die abgesehen von Eigennamen alles klein schreiben wie z.B. Englisch erreichen die besten Erkennungsraten. Sprachen, die Nomen groß schreiben, wie es im Deutschen der Fall ist, erreichen entsprechend schlechtere Ergebnisse. Probleme gibt es bei Sprachen, die keine lateinischen Zeichen verwenden (z.B.: russisch oder chinesisch) und Sprachen, in denen Namen nicht in den Kategorien Vorname und Nachname ausgedrückt werden.

5.4. Beziehungen

Beziehungen werden in den Kategorien quantitativ und qualifiziert angegeben.
quantitative Beziehung sagen aus, wer wen kennt (Bsp. A kennt B). Eine solche Aussage über eine Beziehung zu geben ist relativ einfach. Es bedarf jedoch einer umfangreichen Datenmenge und Zusatzinformationen zu den betroffenen Personen um einen Informationsgewinn zu liefern. Wenn diese Voraussetzungen erfüllt sind, entsteht ein Beziehungskonstrukt. Dieses erlaubt Schlüsse darüber, in welchen Kreisen sich eine Person bewegt.
Die qualifizierte Beziehung beschreibt die Art und Intensität des Verhältnisses von Personen (Bsp. A ist verheiratet mit B; A ist befreundet mit C). Ähnlich wie bei der Georeferenzierung ist es bei den qualifizierten Beziehungen schwierig korrekte Informationen zu erlangen.
Dies liegt zum einen daran, dass die Daten nur mit einer umfangreichen Textanalyse durchgeführt werden können. Dessen Komplexität steigt mit der Detailtiefe der Beziehung. So wird A B niemals mit „Hallo meine Ehefrau B“ ansprechen. Zur Erlangung derartiger Informationen können Daten wie der gemeinsame Nachname in Kombination mit übereinstimmenden Adressdaten dienen.

[2]Exchangeable Image File Format(EXIF) sind Metadaten, die u.a. GPS Daten enthalten können.

Ein weiterer Grund für den erschwerten Umgang mit qualifizierten Beziehungsdaten ist, dass sich diese ändern können. Personen, die beispielsweise Arbeitskollegen waren könnten eine Beziehung eingehen, was zu Fehlern bei Auswertungen führt.
Eine qualifizierte Beziehungsanalyse erfordert einen größeren Aufwand, liefert jedoch bessere Ergebnisse.
Abbildung 1 ein Beziehungsnetzwerk von weltweiten Führungspositionen. Die Daten wurden aus Nachrichtenberichten ermittelt.

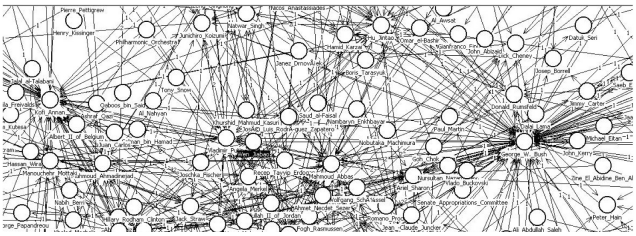

Abbildung 1: Kommunikationen von Führungspersonen auf Basis von Nachrichten aus dem Jahre 2006.[1]

5.5. Ereignisse

Um eine Quelle besser nutzen zu können ist es von Vorteil diese einem Ereignis zuordnen zu können. Hierzu ist es wichtig Ereignisse zu erkennen und diese voneinander zu unterscheiden. Ereignisse mit vielen Quellen werden im allgemeinen als wichtig angesehen. Die Erkennung von Dopplungen, wie in Abschnitt 4.3 beschrieben ist hierbei besonders wichtig.

Zur Erkennung, ob und wie gut ein Dokument in einen Themenbereich passt werden Texte mit Stichworten versehen. Stichwörter sind Wörter mit dem geringsten Informationsgehalt. Das ganze Dokument wird bei diesem Ansatz als Vektor dargestellt. Jedes Stichwort ist eine Dimension in dem Vektor. Die Werte der Dimension ist die Anzahl, die das Wort vorgekommen ist.

Mit dieser Vorgehensweise wird ein Cluster aus Texten erstellt. Das Dokument im Zentrum des Clusters ist das, für das Themengebiet typischste Dokument.

6. KLASSIFIKATOREN

Die automatisierte Analyse von Texten ist immer Fehlerbehaftet. Die Fehleranfälligkeit variiert mit der Art der Analyse, ebenso wie der Schwellenwert an zulässigen Fehlern.
Zur besseren Handhabe mit den verschiedenen Zuständen wurden diese in vier Klassen unterteilt. Diese sind in Tabelle 5 dargestellt.

Tabelle 5: *Klassifikatoren bei der Erkennung*

	ist positiv	ist negativ
positiv klassifiziert	true positive (TP)	false positive (FP)
negativ klassifiziert	false negativ (FN)	true negative (TN)

6.1. Recall (Erinnerung)

Die Erinnerung gibt den Anteil der positiv Objekte an, die positiv klassifiziert wurden. Die Berechnung lautet:

$$\frac{TP}{(TP + FN)}$$

6.2. Accuracy (Treffergenauigkeit)

Die Treffergenauigkeit gibt den Anteil der korrekt Klassifizierten Objekte an. Dieser errechnet sich wie folgt:

$$\frac{TP}{(TP + FP)}$$

6.3. Precision (Präzision)

Die Präzision gibt den Anteil der positiv klassifizierten Objekte an, die tatsächlich positiv sind.

$$\frac{(TP + TN)}{(TP + FP + FN + TN)}$$

7. MENSCHLICHE INTERAKTION

Menschliche Interaktion ist im Gegensatz zur maschinellen Auswertung deutlich kostspieliger und wird deshalb maximal gering gehalten. Abhängig von dem Vorhaben sind jedoch nur gewisse Fehlertoleranzen akzeptabel. Häufig sind besonders FP Ergebnisse ein Problem, da sie zur Verfälschung des Endergebnisses führen.
Um dies zu vermeiden wird der Schwellwert der erreicht werden muss um ein Positives Ergebnis zu liefern erhöht. In der Folge steigen dafür die FN Ergebnisse an. Diese sind für eine korrekte Analyse jedoch wenig Problematisch.
Zur Verbesserung der Ergebnisse und einer besseren Erkennung werden geschulte Personen eingesetzt. Diese überprüfen Ergebnisdaten auf ihre Korrektheit. Häufige Fehler sind z.B. Ereignisse, die jährlich auftreten und deshalb von dem System als eigenständiges und nicht als sich wiederholendes Ereignis erkannt werden.

8. VERTRAUEN UND WAHRHEIT

Das wahrscheinlich problematischste Feld von OSINT ist das Vertrauen in Aussagen und die Prüfung dessen Wahrheitsgehalt.
Bei Kommunikationspartner, die sich gut kennen wird häufig ein Teil der Informationen weg gelassen, da sie beim Kommunikationspartner bekannt sind. Dies tritt insbesondere dann auf, wenn ein Kanalwechsel vollzogen wird. Z.B. wenn eine persönliche Kommunikation auf dem digitalen weg fortgeführt wird.
Ein weiteres Problem bei der Validierung von Aussagen sind Übertreibungen. Diese verfälschen Fakten und eine Korrektur ist nur durch den Abgleich mit Referenzquellen möglich.
Eine ebenso große Herausforderung stellen Unwahrheiten und Ironie dar. Sie sind für technische Systeme nahezu nicht zu erkennen. Ein Mensch kann mit ausreichendem Hintergrundwissen diese unter Umständen ausgleichen.
Ein weiteres Problem sind Vorurteile. Diese sind selbst für erfahrene Analysten ein Problem, da sie das Urteilsvermögen negativ beeinflussen.
Um die Wahrheit einer Quelle zu ermitteln kann diese auf Basis ihrer vergangenen Aussagen eingestuft werden. Quellen mit hoher Seriosität bekommen somit eine höhere Gewichtung, sodass das Ergebnis verbessert werden kann.
Es ist zudem hilfreich Quellen in Kategorien zu unterteilen um deren Aussagen unter Berücksichtigung verschiedener Faktoren zu beurteilen. Mögliche Typen sind:

- unabhängig
- Regierung
- Rebellen
- politisch
- usw.

Zudem wird häufig dem Bericht ein Typ zugewiesen. Mögliche Typen sind:

- erste Hand
- vom hören sagen
- Meinung
- Duplikat
- zensiert
- usw.

9. OSINT WISSENSBASIS

Zur erfolgreichen Durchführung von OSINT bedarf es verschiedener Voraussetzungen. Hierzu zählen unter anderem detaillierte Domain Kenntnisse. Für die automatisierte Analyse ist eine Datenbank mit Metadaten der Domain zu erstellen.
Zur effizienten Nutzung ist es wichtig, Frontend Anwendungen bereit zu stellen. Eine dieser Werkzeuge sind:

- Netminer[6]
- Commetrix[7]
- Gephi[8]
- Cyc[9]

9.1. Länderprofile

Viele Vorkommnisse sind länderübergreifend. Um diese effizient analysieren zu können ist es notwendig jedes Land in Bezug zu seinem Kontext zu betrachten. Hierbei muss die politische, sowie die gesellschaftliche Lage eines Landes beachtet werden. Beispielsweise sind schwere Ausschreitungen in einem Land mit stabiler Wirtschaftslage unwahrscheinlicher und damit von größerer Bedeutung als Ausschreitungen in einem Land in schwieriger Lage.
Die Faktoren, die eine Land, oder eine Region auszeichnen werden als strukturelle Indikatoren bezeichnet.

9.2. Strukturelle Indikatoren

Strukturelle Indikatoren sind qualifizierte sozialpolitische Faktoren. Diese basieren auf verschiedenen statistischen Werten. Einige davon sind:

- Wirtschaftswachstum
- Bildungsgrad
- Lebenserwartung
- Arbeitslosenquote

Um Werte vergleichbar zu halten werden die gleichen Maßstäbe angelegt. So kann zum Beispiel die Pisa Studie[10] zum Vergleich des Bildungsgrats europäischer Länder verwendet werden.
Die strukturellen Indikatoren werden zu Faktoren, wie Konfliktkapazität oder Lebensqualität zusammengefasst. Dies dient letztendlich als Basis für einen direkten Vergleich.

9.3. Dynamische Indikatoren

Dynamische Indikatoren sind Größen, die sich im Gegensatz zu strukturellen Indikatoren häufig ändern. Bestimmte Faktoren sind leicht zu erheben, wie z.B. die Anzahl an Verkehrstoten. Andere Faktoren müssen aus Berichten extrahiert werden, ehe sie genutzt werden können. Diese Werte bilden sich aus der Summe an Vorkommnissen einer Kategorie in einem Bestimmten Zeitrahmen.
Eine Möglichkeit Ereignisse kodiert zu speichern ist das „Integrated Data for Event Analysis (IDEA)“ Framework[11].
Ereignisse können nach Listen, (z.B. von Goldstein[12] aus dem Jahre 1992) bewertet werden. Die Skala reicht von -10 bis 8,3, wobei -10 einem kriegerischen Akt und 8,3 einer starken militärischen Unterstützung gleich kommt. Das Integral über die relevanten Faktoren gibt so den Gefährdungsgrad in einem bestimmten Zeitraum an.
Auf Basis dieser Festen Werte und deren Prognose kann ein Frühwarnsystem entwickelt werden, das versucht zukünftige Krisen hervor zu sagen (siehe Abbildung 2).

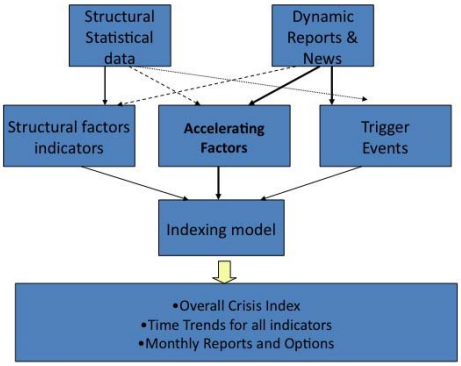

Abbildung 2: Konflikt Früherkennung[1]

9.4. Aufständische und Terroristen

Für das auffinden bestimmter Personen ist das Führen entsprechender Listen von großer Bedeutung. Es ist so einfacher, gesuchte Personen in den gesammelten Intellicence Daten wieder zu finden. Solche Listen werden zumeist händisch mit Daten befüllt, um Fehler auszuschließen. OSINT kann die Informationsgewinnung durch eine Vorfilterung unterstützen.
Das FBI führt z.B. Listen über gesuchte Verbrecher und Terroristen. Diese sind auf der Webseite[13] des FBI einsehbar und können nach belieben durchsucht werden. Mit folgendem Satz wird offen für die Mithilfe an der Ergreifung von Zielpersonen aufgerufen. In manchen Fällen werden zudem Belohnungen für die Ergreifung ausgeschrieben.

> *Protect your family, your local community, and the nation by helping the FBI catch wanted terrorists and fugitives. You can also help reunite missing persons of all ages with their loved ones. Rewards are offered in some cases. Use the Search Center below to find specific cases.*

9.5. Datenbank mit Tupeln

Zur Einordnung von Geschehnissen und der Zuordnung von Aussagen und dessen Gewichtung, ist es notwendig eine hierarchische Liste mit mächtigen Personen aus der Politik und Wirtschaft anzulegen.
Es ist somit möglich Aussagen einer Person in Bezug auf dessen Zugehörigkeit einer Partei oder eines Unternehmens zu interpretieren. Die Hierarchie der Person in seiner Organisation gibt Aussage über Gewichtung einer Aussage.
Die Aussage eines angestellten würde in den meisten Fällen als Einzelmeinung gewertet, wohingegen die Aussage eines Mitglieds des Führungsstabes für die Organisation richtungsweisend sein kann.

10. SPRACHLICHE PROBLEME

In der Arbeit von Watters[2] werden die verschiedenen Probleme der Auswertung von Sprache thematisiert. Ins besonders die Übersetzung führt zu ungewollten Verfälschungen der ursprünglichen Aussage.

10.1. Allgemein

Die Auswertung von Texten über die Sprachgrenzen hinweg ist im allgemeinen ein schwieriges Feld. Durch die Übersetzung von Sprachen gehen Informationen verloren, oder bekommen eine andere Bedeutung. Leicht lässt sich dies durch Dienste wie „Translation Party“[3]nachvollziehen. Ein weiteres Problem stellt die Wortreihenfolge in unterschiedlichen Sprachen dar.
In speziellen Bereichen ist eine Analyse jedoch erheblich einfacher. Dies ist insbesondere dann der Fall, wenn der Themenkomplex durch eine konkrete Sprache geprägt ist. Beispiele sind die Informatik, die größtenteils englische Begriffe verwendet, sowie die Medizin, die in gleichem Maße auf die lateinische Sprache zurück greift.

10.2. Metaphern, Gleichnisse und Sprichworte

Wie die Analyse von Watters zeigt ist insbesondere die Auswertung von Metaphern, Gleichnissen und Sprichworten ein Problem. Alle drei Begriffsgattungen basieren auf einer nahezu exakten Wiedergabe und selbst die Nutzung von Synonymen macht eine Metapher oder ein Gleichnis unverständlich. Zudem ergibt eine solche Wortkonstruktion selbst korrekt übersetzt in der übersetzten Form kein Sinn ergeben. Folgendes Beispiel verdeutlicht dies.
„Jemand ist noch grün hinter den Ohren“ bedeutet grammatikalisch korrekt übersetzt „Someone is still green behind the ears“. Dennoch würde ein englischer Muttersprachler das Sprichwort nicht erkennen. Korrekt übersetzt müsste das Sprichwort „Someone is still wet behind the ears“ oder „Someone is half-baked“ lauten.

[3] `http://translationparty.com/` übersetzt einen englischen Ausdruck ins japanische und zurück, bis sich das Ergebnis nicht mehr ändert.

10.3. direkte(wörtliche) Übersetzung

Die wörtliche Übersetzung ist die einfachste Art Begriffe in eine andere Sprache zu übersetzen, da hierbei keine Grammatikkenntnisse vorhanden sein müssen. Dies ermöglicht eine besonders einfache Behandlung unter der Nutzung von Wörterbüchern. Geeignet ist diese Art der Übersetzung nur für einzelne Wörter.
Durch die Unabhängigkeit zum Kontext ist diese Art der Übersetzung nicht für ganze Sätze geeignet und selbst bei einzelnen Worten gibt es Probleme. Das Englische Wort „fast“ kann mit „schnell“ übersetzt werden. Schnell kann wiederum mit „quick“ übersetzt werden, wodurch die Bedeutung verfälscht wird.

10.4. Kontextsensitive Übersetzung

Im Gegensatz zu der direkten Übersetzung ist die kontextsensitive Übersetzung erheblich komplexer. Dieses Vorgehen ist jedoch, wie bereits erwähnt, notwendig um Phrasen korrekt zu übersetzen. Hierzu werden zwei primäre Ansätze verfolgt.
Eine Möglichkeit ist es, korrekt übersetzte Phrasen zu hinterlegen. Dies kann ein Großteil, der in Abschnitt 10.2 und 10.3 beschriebene Probleme lösen. Der Ansatz nutzt Transferfunktion und erfordert die folgenden drei Schritte:

- Analyse der Eingabephrase.
- Entwicklung einer Transferfunktion.
- Erstellung einer Darstellung in der Zielsprache.

Dieser Ansatz hat die Schwäche, dass jede Sprache in jede andere übersetzt werden muss.
Ein weiterer Möglichkeit ist ein interlinguale Ansatz. Hierbei wird von Sprachen in eine Metasprache abstrahiert. und von dort aus in die Zielsprache übersetzt. Bei diesem Vorgehen ergeben sich lediglich zwei Schritte:

- Übersetzung in die Metasprache
- Übersetzung von der Metasprache in die Zielsprache

Der Interlinguale Ansatz bietet den Vorteil, dass jede Sprache nur einmal in die Metasprache übersetzt werden muss und von dort aus in die Zielsprache.
Beide genannte Verfahren sind ineffizienter als die direkte Übersetzung. In der Regel spielt der Faktor, der Korrektheit eine übergeordnete Rolle und die Effizienz ist ein Problem von abnehmender Bedeutung. Dies liegt an der Entwicklung effizienterer Algorithmen und der Relativierung durch „Moore's law“[14].

11. EINSATZGEBIETE VON OSINT

Open Source Intelligence ist ein komplexes Feld, dass von einer Vielzahl von Faktoren abhängt. In den vergangenen Jahren hat dieser Bereich der Datenauswertung, auf Grund von gestiegener Verfügbarkeit von freien Daten, massiv an Popularität gewonnen. OSINT ist zu einem der wichtigsten Informationsquellen für Regierungen, Geheimdienste und Konzernen geworden.
Die Gruppen, die OSINT betreiben verfolgen unterschiedliche Ziele.
Regierungen sind an der besseren Organisation des Öffentlichen Lebens interessiert. In der Arbeit von Prasad et al.[4] wird OSINT für die Bedarfsanalyse für den Bau von neuen Schulen verwendet. Regierungen können mit OSINT zudem den Erfolg von Gesetzesvorhaben, sowie die Popularität von politischen Maßnahmen messen.
Geheimdienste Nutzen OSINT für die Verfolgung von Personen im In- und Ausland. Der Bundesnachrichtendienst(BND) besitzt beispielsweise eine eigene Abteilung, die sich ausschließlich mit OSINT beschäftigt. Laut offizieller Webseite[15] macht OSINT quantitativ den größten Teil der Informationsgewinnung aus.
Firmen nutzen OSINT für verschiedene Zwecke. Es werden Informationen über Mitbewerbern, dem Markt und Kunden gesammelt. Detaillierte Informationen über den Markt und Mitbewerbern erleichtert die Einführung neuer Produkte. Informationen über Kunden unterstützt das Werben neuer Kunden.

12. ZUSAMMENFASSUNG

Die Informationsgewinnung in einer Zeit, der omnipräsenten Präsenz, des Web 2.0 stellt fundamental andere Aufgaben and Datenanalysten. Die Frage, wie und woher Daten genommen werden, ist der Frage gewichen, wie die gewünschten Informationen aus Flut an Informationen extrahiert werden können.
Für die Zukunft müssen Werkzeuge entwickelt werden, die Fehlerquellen weiter reduzieren und eine bessere interdisziplinäre Auswertung erlauben. Hierzu ist der verstärkte Einsatz von neuronalen Netzen denkbar.
Dennoch werden die Werkzeuge vermutlich niemals perfekt sein, wodurch menschliche Interaktion immer nötig sein wird. Der Fokus neuer Werkzeuge muss deshalb darauf liegen, Analysten bestmöglich bei ihrer Arbeit zu unterstützen.
Die Tatsache, dass Regierungen, Geheimdienste und Firmen, Daten im öffentlich zugänglichen Bereich des Internets, nahezu uneingeschränkt mitlesen und auswerten, hat Auswirkungen auf unsere Gesellschaft. Gerade ältere Menschen und Personen aus weniger technischen Branchen sind sich dieser Auswirkung oft nicht bewusst.
Der Autor dieser Arbeit hält daher Aufklärung und einen gesellschaftlichen Diskurs über den Umgang mit solchen technischen Systeme für unvermeidlich. Anderenfalls droht der Verlust jegliche Privatsphäre im Rausch des Fortschritts.

13. REFERENCES

[1] C. Best, “Challenges in open source intelligence,” in *Intelligence and Security Informatics Conference (EISIC), 2011 European*, Sept 2011, pp. 58–62.

[2] P.A. Watters, “Challenges to automated allegory resolution in open source intelligence,” in *Cybercrime and Trustworthy Computing Workshop (CTC), 2012 Third*, Oct 2012, pp. 14–18.

[3] P. Eachus and B. Short, “Development of a hybrid decision support system for intelligence analysis,” in *Computing and Convergence Technology (ICCCT), 2012 7th International Conference on*, Dec 2012, pp. 745–748.

[4] K.H. Prasad, S. Soni, T.A. Faruquie, and L.V. Subramaniam, “Data consolidation solution for internal security needs,” in *Service Operations and Logistics, and Informatics (SOLI), 2012 IEEE International Conference on*, July 2012, pp. 84–89.

[5] B. Sterling, *Zeitgeist*, Bantam Spectra Books. Bantam Books, 2001.

[6] Cyram, “Netminer,” `http://netminer.com/index.php`, 2001, [Online; aufgerufen am 17.02.2015].

[7] Matthias Trier, "Commetrix," http://www.commetrix.de/, 2008, [Online; aufgerufen am 17.02.2015].

[8] Mathieu Bastian, "Gephi," https://gephi.github.io/, 2008, [Online; aufgerufen am 17.02.2015].

[9] Cycorp, "Cyc," http://www.cyc.com/, 1984, [Online; aufgerufen am 17.02.2015].

[10] OECD, "Pisa - internationale schulleistungsstudie der oecd," http://www.oecd.org/berlin/themen/pisa-internationaleschulleistungsstudiederoecd.htm, 2012, [Online; aufgerufen am 17.02.2015].

[11] The IDEA Project, "Integrated data for event analysis (idea)," http://vranet.com/IDEA.aspx, 1998-2014, [Online; aufgerufen am 18.02.2015].

[12] Joshua S. Goldstein, "A conflict-cooperation scale for weis events data," *The Journal of Conflict Resolution*, vol. 36, no. 2, pp. 369–385, jun 1992.

[13] FBI, "Wanted list," http://www.fbi.gov/wanted, 2015, [Online; aufgerufen am 18.02.2015].

[14] Gordon E Moore et al., "Cramming more components onto integrated circuits," *Proceedings of the IEEE*, vol. 86, no. 1, pp. 82–85, 1998.

[15] BND, "Osint," http://www.bnd.bund.de/DE/Arbeitsfelder/Informationsgewinnung/OSINT/osint_node.html, 2015, [Online; aufgerufen am 19.02.2015].